DNA Memory with 16.8M Addresses

Masahito Yamamoto[1,2], Satoshi Kashiwamura[3], and Azuma Ohuchi[1,2]

[1] Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan
[2] CREST, Japan Science and Technology Agency (JST), Japan
[3] HITACHI, Co. Ltd., Japan

Abstract. A DNA Memory with over 10 million (16.8M) addresses was achieved. The data embedded into a unique address was correctly extracted through addressing processes based on the nested PCR. The limitation of the scaling-up of the proposed DNA memory is discussed by using a theoretical model based on combinatorial optimization with some experimental restrictions. The results reveal that the size of the address space of the DNA memory presented here may be close to the theoretical limit. The high-capacity DNA memory can be also used in cryptography (steganography) or DNA ink.

Keywords: DNA Computing, DNA memory, NPMM, Theoretical capacity.

1 Introduction

Deoxyribonucleic acid (DNA) is well known as the blueprint of life, while it is also an attractive material because of its excellent properties such as minute size, extraordinary information density, and self-assembly. Focusing on these facts in recent years, various works, especially studies in the research field of DNA computing, have tried to develop a method for solving the combinatorial problems or for developing DNA machines such as a DNA automata [1][2][3][4].

One of the most promising applications of DNA computing might be a DNA memory. DNA molecules can store a huge amount of information in their sequences in extremely small spaces. The stored information on DNA can be kept without deteriorating for a long period of time because DNA is very hard to collapse [5]. Baum was the first to propose DNA memory, which can have a capacity greater than the human brain in minute scale [6]. The model enables a massively parallel associative search in a vast memory by utilizing the parallelism of DNA's hybridization. Rife et al. and Neel et al. have described DNA memory similar to that of Baum's model [7][8]. Recently, Chen et al. have proposed a DNA memory model that is capable of learning new data and recalling data associatively [9]. Although various research has been conducted on the construction of DNA memory, almost all of the works involve only proposals of models or only the preliminary experiments on a very small scale. Even if they could operate correctly on a small scale, it is doubtful that the operation would be successful in larger DNA memory because the efficiency and specificity of DNA's chemical

M.H. Garzon and H. Yan (Eds.): DNA 13, LNCS 4848, pp. 99–108, 2008.

reaction become much more severe. Therefore, it is very important to prove the technology through actual demonstration of the construction and addressing of DNA memory.

In this work, a DNA memory with 16.8M addresses is achieved. Our proposed DNA memory is addressable by using nested PCR and is named Nested Primer Molecular Memory (NPMM) [10][11][12]. The size of NPMM may be the largest pool of DNA molecules, which means that there are a large number of kinds of DNA sequences and any kind of DNA sequence can be extracted from the pool. The advantage of our memory is addressing based on amplification, which can amplify the target sequences and not amplify the non-target sequences. By using this amplification in several addressing steps, the probability of extraction of non-target DNA sequences can be very low. In fact, it is shown that any data can be retrieved with very high fidelity. The limitation of scaling-up of the proposed DNA memory is also discussed by using a theoretical model based on combinatorial optimization with some experimental restrictions. The results reveal that the size of the address space of the DNA memory presented here may be close to the theoretical limit. The high-capacity DNA memory can be also used in cryptography (steganography) or DNA ink [13][14][15][16][17].

2 Nested Primer Molecular Memory

NPMM is the pool of DNA strands such that each strand codes both data information and its address information. The data information (ex. binary data, strings, etc.) is expressed by encoded base sequence. The address information consists of several layers and each layer contains several components (specific DNA sequences) and is expressed by the combination of components on each layer. These layers are divided into two portions and are located on both sides of the data. In this work, we deal with the following one called 16.8M-NPMM: three layers on each side (named XY, $X \in \{A, B, C\}$, $Y \in \{L, R\}$) and sixteen sequences (20 mer) on each layer (named XYi, $i \in \{0, 1, 2 \dots 15\}$). Each DNA molecule is structured such as $CL*$-$CLlink$-$BL*$-$BLlink$-$AL*$-$ALlink$-$Data$-$ARlink$-$AR*$-$BRlink$-$BR*$-$CRlink$-$CR*$ as shown in Fig 1. The notation '-' means the concatenation of DNA sequences. The $XYlink$ ($X \in \{A, B, C\}$, $Y \in \{L, R\}$) indicates a linker section (20 mer) and these are used to construct 16.8M-NPMM. The address information is expressed by the combination of XYi denoted by such as [CLi, BLj, ALk, ARl, BRm, CRn] $(i, j, k, l, m, n \in \{0, 1, \dots 15\})$. The address space of 16.8M-NPMM is about 16.8 million $(= 16^6 = 16,777,216)$.

Operations for retrieving the stored data are executed by specifying each address layer based on PCR. For easy understanding, we will now explain how to retrieve the target information stored at [$CL0$, $BL2$, $AL4$, $AR5$, $BR3$, $CR1$] from 16.8M-NPMM (Fig. 2). For the first operation, PCR is performed for 16.8M-NPMM using $CL0$ and $\overline{CR1}$ as primer pairs ($\overline{x}$ is the complementary DNA

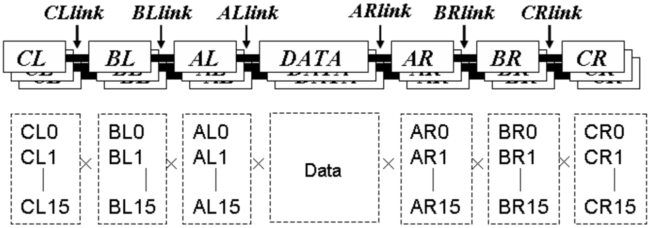

Fig. 1. Sequence structure of each DNA strand in 16.8M-NPMM. The area expressing the address information consists of six layers (XY, $X \in \{A, B, C\}$, $Y \in \{L, R\}$) and ten sequences are defined in each layer (named XYi, $i \in \{0, 1, 2 \ldots 15\}$). The address space of 16.8M-NPMM is over 10 million.

sequence of x). As a result, we can extract the collection of DNA molecules containing $CL0$ and $CR1$ from 16.8M-NPMM and exclude all DNA molecules without $CL0$ or $CR1$. This is because PCR yields a significant difference in the concentration between the amplified and non-amplified DNA molecules; therefore, we can disregard the non-amplified DNA. Next, we perform the second PCR using $BL2$ and $\overline{BR3}$ for the diluted solution after the first PCR. At this point, we can extract the DNA molecules containing $CR0$, $CR1$, $BL2$ and $\overline{BR3}$. Next, we perform the third PCR for each diluted solution using $AL4$ and $\overline{AR5}$. After all PCRs are completed, we can extract only the DNA molecule expressing [$CL0$, $BL2$, $AL4$, $AR5$, $BR3$, $CR1$] that codes the target data. Sequencing and decoding the extracted DNA molecules allows us to retrieve the target data.

3 Construction and Addressing of 16.8M-NPMM

We carried out laboratory experiments to verify the behavior of NPMM with over ten million address spaces. For simplicity and easy detection, the stored information in 16.8M-NPMM is either Data20 (20 mer), Data40 (40 mer) or Data60 (60 mer). Data40 is embedded into a unique address $[CL0, BL0, AL0, AR0, BR0, CR0]$, and Data60 is embedded into another unique address $[CL8, BL8, AL8, AR8, BR8, CR8]$, while Data20 is in all addresses. By observing the results of data extraction, the success of the addressing is evaluated. Note that Data40 and Data60 are stored in only one address among 16.8M addresses. We can detect the behavior very easily using only gel electrophoresis due to the difference in length. All DNA sequences in this work were designed by using a Two-Step Search Algorithm to avoid any mis-hybridization based on Hamming Distance [18]. Of course, another algorithm for designing DNA sequences can be used[19][20][21]. Moreover, to avoid a secondary structure, they are designed so that the free energy of each DNA molecule can reach a high score by using an *m-fold* algorithm customized for DNA molecules [22][23][24][25].

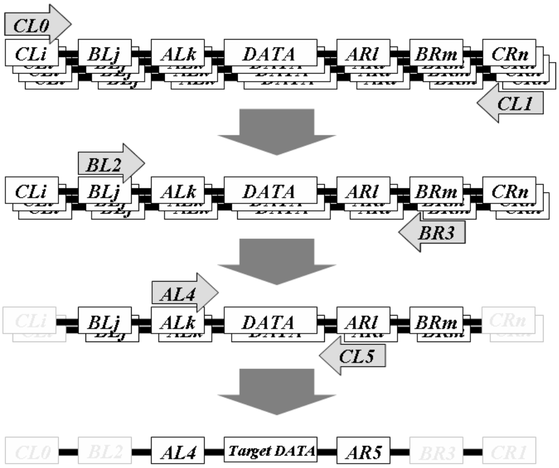

Fig. 2. Operations for retrieving the target DNA from NPMM are implemented in the nested PCR. In this case, the target address is [$CL0$, $BL2$, $AL4$, $AR5$, $BR3$, $CR1$]. Only if specifying the target address is completed will the target DNA be extracted from the large mixture of DNA molecules. If any other address is specified, non-target DNA will be extracted. The faint portions are the areas eliminated by the previous PCR.

3.1 Construction of 16.8M-NPMM

We prepare a DNA molecule such as Ci-$CLlink$, $CLlink$-Bj-$BLlink$, $BLlink$-Ak-$ALlink$, $ALlink$-$Data20$-$\overline{ARlink}$, $ALlink$-$Data40$-$\overline{ARlink}$, $BRlink$-ARl - $ARlink$, $CRlink$-BRm-$BRlink$ and CRn-$CRlink$ $(i, j, k, l, m, n \in \{0, 1 \ldots 15\})$. The first step is to perform PCR for DNA containing $Data20$ using $BLlink$-Ak-$ALlink$ and $BRlink$-ARl-$ARlink$ as primer pairs. This leads the AL and AR layers to Data20 based on the priming reaction of the linker sections. Similarly, other layers are also integrated using linker sections as knots. These steps produced about 16.8M whole addresses. Specific data (Data40 and Data60) are also created using the same method. There were no unwanted products through each experimental process (data not shown). Next, we measure the total amount of the whole pool of memory molecules and the Data40 and Data60 are mixed into the pool so that the number of each longer DNA and other DNA are equivalent.

3.2 Addressing from 16.8M-NPMM

We used sixteen kinds of addresses shown in Table 1 as test samples. The image of PolyAcrylamide Gel (Fig. 3) shows the results of laboratory experiments for the addressing operations. The three lanes in each surrounded area correspond to each addressing. The most right lane indicates the result after the final PCR. From the results of Fig. 3, we can see there was no minor band throughout the whole experiment. DNA product 120 bp long was obtained only in address All 0

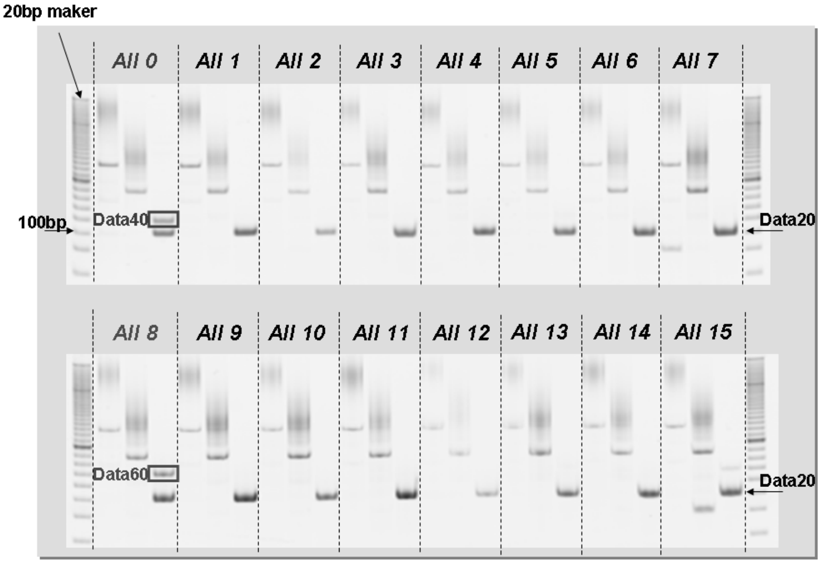

Fig. 3. Result of PolyAcrylamide Gel Electrophoresis for addressing: he three lanes in each surrounded area correspond to each addressing. Three lanes in each surrounded area correspond to the result of the first PCR, the second PCR and the final PCR, respectively. All$i$ means the address [CLi, BLi, ALi, ARi, BRi, CRi].

Table 1. Addressing Samples

Label	Address	Data
All 0	$[CL0, BR0, AL0, AR0, BL0, CR0]$	Data20, Data40
All 8	$[CL8, BL8, AL8, AR8, BR8, CR8]$	Data20, Data60
All i $(i = 0, \cdots, 15, i \neq 0, 8)$	$[CLi, BLi, ALi, ARi, BRi, CRi]$	Data20

and DNA product 140 bp long was obtained only in address All 8, while all other addresses output product 100 bp long. Therefore, we can successfully extract the corresponding data of each address among over 10 million addresses. Based on these facts, although we did not confirm all addresses, we assume that NPMM could reliably screen and extract a target DNA molecule from over 10 million DNA mixtures. In DNA memory, it is most important to extract the target data with high fidelity. Therefore, this experiment strongly supports the effectiveness of NPMM.

4 Theoretical Analysis of Capacity Limitation

It is obvious that the scaling up of NPMM is limited due to some physical or chemical restrictions such as the limitation of the amount of DNA in solution

$$
\begin{array}{lll}
\mathit{maximize} & \displaystyle\prod_{i=1}^{L} n_i^2 & \\
\mathit{sub.to} & n_i \geq 2 \ (i \in \{1, 2, \dots L\}) & (1) \\
 & N \geq \displaystyle\sum_{i=1}^{L} 2n_i & (2) \\
 & v(2^{c_i} - 1) \displaystyle\prod_{j=i+1}^{L} n_j^2 + 2vc_i(n_i - 1) \prod_{j=i+1}^{L} n_j^2 \leq P \qquad \forall i \in \{1, 2, \dots L\} & (3) \\
 & amp_{c_i} > \alpha \times non\_amp_{c_i} \qquad \forall i \in \{1, 2, \dots L\} & (4) \\
 & 0 \leq c_i \leq max\_cycle \qquad \forall i \in \{1, 2, \dots L\} & (5) \\
 & P,\ L,\ N,\ v,\ \alpha,\ max\_cycle : \text{integer (given)} &
\end{array}
$$

Fig. 4. Combinatorial optimization problem for NPMM analysis

and the number of DNA sequences available for the layer area. Therefore, it is important to discern how far the capacity of NPMM can be extended. We propose an addressing model for analyzing NPMM capacity limitations based on the combinatorial optimization problem. By solving this optimization problem, the maximum capacity of NPMM and the optimal assignment of address sequences to each address area can be obtained.

The capacity of NPMM depends on the address space and size of each datum. However, since enlarging address space is more difficult experimentally, we mainly discuss the enlargement of address space here. We deal with such models as L layers (L is integer grater than one) and n_i DNA sequences located on the ith layer ($n_i \geq 2, i \in \{1, 2, \dots L\}$, i indicates a layer location from left side). For simplicity, the data area is abbreviated.

First, we consider the requirements for operating NPMM correctly and then write out a mathematical formula as constraints. Then, based on the constraints, we establish an expression to calculate the maximum capacity of NPMM (Fig. 4). The expression outputs the arrangements of n_i ($i \in \{1, \dots L\}$) maximizing the address space while satisfying the constraints when input parameters are given. The input parameters should be the ones used in laboratory experiments to reflect the actual environment.

4.1 Model of the Addressing of NPMM

Input Parameters: P is the total number of primer molecules and should range from 1.2×10^{13} to 5.0×10^{13}. These values are practically appropriate for $100\mu l$

of PCR mixture. L is the number of layers. N is the total number of DNA sequences for address layers, which are designed so carefully that they can be available as PCR primer. v expresses the initial amount of each DNA strand in NPMM and v must be greater than one. Otherwise, it means that several address units are missing in memory. α expresses the rate between the total number of amplified DNA strands and that of non-amplified DNA strands after PCR. To succeed with each addressing process, α should be large enough so that the non-amplified DNA strands can be eliminated. max_cycle is the upper limit of PCR cycles.

Objective function: The target function is calculated as follows:

$$Capacity = \prod_{i=1}^{L} n_i^2 .$$

$Capacity$, which is the width of address-space, depends on n_i. That is, this problem is an optimization problem to explore an arrangement n_i to maximize $Capacity$ while satisfying the following constraints.

Constraints: These constraints are provided to ensure NPMM's behavior. Constraint (1) is obviously to express the address structure of NPMM. Constraint (2) ensures that all sequences located on each layer must be well-designed DNA sequences that can avoid any mis-hybridization or mis-priming. Constraints (3), (4) and (5) are established to make each addressing operation possible physically, and these three constraints must be satisfied in each ith addressing operation ($i \in \{1, 2, \ldots L\}$). c_i is the number of PCR cycles for addressing the ith layer. Constraint (3) ensures that each addressing operation is able to work. The terms amp_{c_i} and $non\_amp_{c_i}$ are calculated as follows:

$$\begin{aligned} amp_{c_i} &= v \prod_{j=i+1}^{L} n_j^2 + v(2^{c_i} - 1) \prod_{j=i+1}^{L} n_j^2 \\ non\_amp_{c_i} &= v(n_i^2 - 1) \prod_{j=i+1}^{L} n_j^2 + 2vc_i(n_i - 1) \prod_{j=i+1}^{L} n_j^2 . \end{aligned}$$

amp_{c_i} and $non\_amp_{c_i}$ indicate the total amount of amplified and non-amplified DNA strands after the ith PCR, respectively. Therefore, amp_{c_i} must greatly exceed $non\_amp_{c_i}$[26][27]. (Here, we use α.) Constraint (4) expresses the total amount of amplified DNA molecules in the ith PCR. According to the principle of PCR, the total amount of amplified DNA in the PCR is never greater than that of the PCR primer. The former term is the increment of target DNA strands, which are amplified exponentially. The latter term is the increment of non-target DNA strands, which are linearly amplified from only one side priming. Constraint (5) is defined to avoid excess PCR cycles because excess PCR cycles cause unwanted reaction. That is, constraints (3), (4) and (5) ensure that a large difference in concentration is acquired between amplified and non-amplified DNA after PCR for addressing the ith layer.

4.2 Computational Result

According to this expression, we explore the theoretical maximum capacity of NPMM by exhaustive search. The input parameters are shown in Table 2. The computational results are shown in Table 3.

Table 2. Parameter settings

P	5.0×10^{13}
L	2,3,4
N	200
α	1000
v	100
max_cycle	30

P is set to the standard number of primers used in 100 μl of PCR mixture. L is the very important parameter for defining capacity. However, a large L makes the laboratory experiments cumbersome and complicated. In this paper, we selected $L = 2, 3, 4$ and analyzed in these cases. As for α and v, what values are appropriate for a successful addressing operation are as yet unclear. Therefore, we negatively set these values to 1,000 and 100, respectively. N is determined based on works in the research field of DNA word design problems.

Table 3. Theoretical capacity of case of $L = 2, 3, 4$. The numbers in brackets means the assignment of the number of address sequences, In the case of $L = 2$, it is shown that the optimal assignment of the address sequences to (BL, AL, AR, BR) is (50, 50, 50, 50).

Layer	L=2	L=3	L=4
Capacity	6,250,000 (50,50,50,50)	274,233,600 (69,16,15,15,16,69)	297,666,009 (71,9,9,3,3,9,9,71)

A chemical reaction inevitably includes fluctuations (for example, the deviation of the number of DNA strands in NPMM and that of the amplification efficiency), so v should be greater than 100 to accomplish operations of NPMM with a high degree of fidelity. Therefore, the limitation of NPMM is expected up to MEGA order. However, this is an ideal one when a PCR reaction is carried at the maximum efficiency (amplification efficiency is always twice, and the unwanted reaction does not occur). Probably, practicable marginal capacity will be slightly smaller, and so the 16.8M-NPMM of $L = 3$ we constructed here has a large capacity considerably close to the practical limit.

5 Concluding Remarks

In this work, we dealt with NPMM, which is our proposed DNA memory. We constructed NPMM with over 10 million address spaces (16.8M-NPMM), and then several addresses were operated. From the experimental results, they showed completely correct behavior. Furthermore, since we showed that 16.8M-NPMM works with very high fidelity, we can conclude that DNA memory with 16.8M address spaces is achieved. We established a technology that selects only specific DNA from many kinds of DNA in mixture. The achievement of 16.8M DNA memory may be the largest pool of DNA mixture so far.

The latter part mainly discusses the theoretical limitations of NPMM's capacity. The behavior of NPMM was expressed by mathematical formula. By solving a combinatorial optimization problem, we could estimate that the theoretical limitation is MEGA order. Taking the efficiency of a chemical reaction into consideration, the process will become less practical.

Acknowledgements

We thank M. Hagiya, A. Suyama, A. Kameda and S. Yaegashi for helpful advice and discussions. The work presented in this paper was partially supported by a Grant-in-Aid for Scientific Research on Priority Area No.14085201 and a Grant-in-Aid for Young Scientists (A) No. 17680025, Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Adleman, L.M.: Molecular Computation of Solutions to Combinatorial Problems. Science 266, 1021–1024 (1994)
2. Lipton, R.: DNA solution of hard combinatorial problems. Science 268, 542–545 (1995)
3. Braich, R.S., Chelyapov, N., Johnson, C., Rothemund, P.W.K., Adleman, L.: Solution of a 20-Variable 3-SAT Problem on a DNA Computer. Science 296, 499–502 (2002)
4. Benenson, Y., Paz-Elizur, T., Adar, R., Keinan, E., Livneh, Z., Shapiro, E.: Programmable and autonomous computing machine made of biomolecules. Nature 414, 430–434 (2001)
5. Wong, P.C., Wong, K.K., Foote, H.: Organic Data Memory Using the DNA Approach. Communications of the ACM 46(1), 95–98 (2003)
6. Baum, E.B.: Building an Associative Memory Vastly Larger Than the Brain. Science 268, 583–585 (1995)
7. Reif, J.H., LaBean, T.H., Pirrung, M., Rana, V.S., Guo, B., Kingsford, C., Wickham, G.S.: Experimental Construction of Very Large Scale DNA Databases with Associative Search Capability. In: Jonoska, N., Seeman, N.C. (eds.) DNA Computing. LNCS, vol. 2340, pp. 231–247. Springer, Heidelberg (2002)
8. Neel, A., Garzon, M.H., Penumatsa, P.: Improving the Quality of Semantic Retrieval in DNA-Based Memories with Learning. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3213, pp. 18–24. Springer, Heidelberg (2004)

9. Chen, J., Deaton, R., Wang, Y.Z.: A DNA-based memory with in vitro learning and associative recall. Natural Computing 4(2), 83–101 (2005)
10. Kashiwamura, S., Yamamoto, M., Kameda, A., Shiba, T., Ohuchi, A., Hierarchical, D.N.A.: Memory based on Nested PCR. In: Hagiya, M., Ohuchi, A. (eds.) DNA Computing. LNCS, vol. 2568, pp. 112–123. Springer, Heidelberg (2003)
11. Kashiwamura, S., Yamamoto, M., Kameda, A., Shiba, T., Ohuchi, A.: Potential for enlarging DNA memory: The validity of experimental operations of scaled-up nested primer molecular memory. BioSystems 80, 99–112 (2005)
12. Kashiwamura, S., Yamamoto, M., Kameda, A., Ohuchi, A.: Experimental Challenge of Scaled-up Hierarchical DNA Memory Expressing a 10,000-Address Space. In: Preliminary Proceeding of 11th International Meeting on DNA Based Computers. vol. 396 (2005)
13. Clelland, C.T., Risca, V., Bancroft, C.: Hiding message in DNA microdots. Nature 399, 533–544 (1999)
14. Hashiyada, M.: Development of Biometric DNA Ink for Authentication Security. Tohoku J. Exp. Med. 204, 109–117 (2004)
15. Hashiyada, M., Itakura, Y., Nagashima, T., Nata, M., Funayama, M.: Polymorphism of 17 STRs by multiplex analysis in Japanese population. Forensic Sci. Int. 133, 250–253 (2003)
16. Itakura, Y., Hashiyada, M., Nagashima, T., Tsuji, S.: Proposal on Personal Identifiers Generated from the STR Information of DNA. Int. J. Information Security 1, 149–160 (2002)
17. Kameda, A., Kashiwamura, S., Yamamoto, M., Ohuchi, A., Hagiya, M.: Combining randomness and a high-capacity DNA memory. DNA13 (submitted 2007)
18. Kashiwamura, S., Kameda, A., Yamamoto, M., Ohuchi, A.: Two-Step Search for DNA Sequence Design. IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences E87-A (6), 1446–1453 (2004)
19. Deaton, R., Murphy, R.C., Garzon, M., Franceschetti, D.R., Stevens Jr, S.E.: Good Encoding for DNA-Based Solutions to Combinatorial Problems. In: Landweber, L.F., Baum, E.B. (eds.) DNA Based Computers II. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, pp. 247–258 (1999)
20. Tanaka, F., Kameda, A., Yamamoto, M., Ohuchi, A.: Design of nucleic acid sequences for DNA computing based on a thermodynamic approach. Nucleic Acids Research 33, 903–911 (2005)
21. Tulpan, D.C., Hoos, H.H., Condon, A.: Stochastic Local Search Algorithms for DNA word Design. In: Hagiya, M., Ohuchi, A. (eds.) DNA Computing. LNCS, vol. 2568, pp. 229–241. Springer, Heidelberg (2003)
22. Lyngso, L.B., Zuker, M., Pedersen, C.N.: Fast evaluation of internal loops in RNA secondary structure prediction. Bioinformatics 15, 440–445 (1999)
23. SantaLucia, J., Allawi, H.T., Seneviratne, P.A.: Improved nearest-neighbor parameters for predicting DNA duplex stability. Biochemistry 35, 3555–3562 (1996)
24. Sugimoto, N., Nakano, S., Yoneyama, M., Honda, K.: Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. Nucleic Acids Research 24, 4501–4505 (1996)
25. Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research 9, 133–148 (1981)
26. McPherson, M.J., Hames, B.D., Taylor, G.R.: PCR A Practical Approach. IRL Press (1995)
27. McPherson, M.J., Hames, B.D., Taylor, G.R.: PCR2 A Practical Approach. IRL Press (1993)